INEXACT MATCHING OF PROPER NAMES IN SINHALA

M.SC. IN COMPUTER SCIENCE

S. C. FERNANDO

UNIVERSITY OF MORATUWA, SRI LANKA

DECEMBER 2007

INEXACT MATCHING OF PROPER NAMES IN SINHALA

S. C. FERNANDO

This dissertation was submitted to the
Department of Computer Science and Engineering
of the University of Moratuwa
in partial fulfilment of the requirements for the
Degree of M.Sc. in Computer Science
specializing in Software Architecture

Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
December 2007

DECLARATION

I, S. C. Fernando hereby declare that the work included in this dissertation in part or whole has not been submitted for any other academic qualification at any institution.

Prof. Gihan Dias S. C. Fernando

Supervisor

ABSTRACT

With the advancement of technology, the need for maintaining national data and information becomes important. Most of these data and information have to be maintained in the local languages because majority of the Sri Lankans are still not very conversant in English. Therefore when public organizations embrace IT, their data including personal data has to be maintained in local languages. When data and information are available in the local language, searching and retrieving them using the local language become essential.

Proper nouns have an inherent problem because a given proper noun, for example a name can be spelt in several different ways. This problem becomes more prominent when a name from one language origin is spelt using another language. For example, the Sinhala name විශාඛා can be spelt in several ways such as විසාකා, විසාඛා or විශාකා using Sinhala itself. Therefore, one who would search an information store for a proper name may not encounter a match, if a different spelling is used to search from that being stored.

This research was to provide a solution to the problem mentioned above using Sinhala language. That is to build a rule based search application that would take a Sinhala input string, search an information store and retrieve matching results even if they were stored with a different spelling.

This was achieved by building a rule base to replace characters of a key word with different characters in order to generate a set of words with different spelling. Then this set of words is searched in the information store and results are displayed. Rules were organized in different levels so that the user can select the level of character replacement, thus it would retrieve matches with a slight spelling difference or retrieve matches with drastic spelling differences. A special rule set was built for matching Tamil names written in Sinhala. The user has option to independently enable/disable this rule set. An application, which uses a general-purpose rule engine to process rules was designed and implemented to demonstrate this technology. This application consist of a web based user interface and a sample database as the information store. This was designed in a layered architecture such that future expansions and component reuse can be done. All character replacement rules are declared in text files, so changes and updates to the rule base can be done without modifying the system.

It is shown that the application, with the rule base that was built, will provide a solution to the proper name search problem stated above. This system can be integrated with future information systems in government and business organisations.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who helped me to successfully complete the M. Sc. program and this research project.

First and foremost, I offer my deepest gratitude to my supervisor Prof. Gihan Dias, who envisioned this research idea. Despite his busy schedules, he persistently spent adequate time to review the progress, guide and direct me throughout the research period. His advice regarding various aspects of multilingual computing was invaluable in completing this research project successfully.

I would also like to thank Dr. Sanath Jayasena, my co-supervisor for his continual reviews that helped a lot in completing the research and this dissertation in a timely manner. The weekly activities and progress reviews organized by him helped to keep my focus on the research.

I am grateful to the other staff members of the Department of Computer Science and Engineering and visiting lecturers who had given useful advice and direction either during progress reviews or during lectures. Contents in some of the course modules in the M. Sc. programme were directly relevant to this research and the dissertation. My thanks go to all the lecturers who conducted those modules.

My sincere thanks go to the researchers in LTRL of UCSC who had published valuable resources relevant to multilingual computing in their site. Some of these resources were instrumental in completing this research project. I also thank the past researchers and authors of materials that I reviewed for my research.

I am indebted to the open source community for providing powerful software tools, documentations and submitting forum posts, which were crucial in completing this research project successfully. JBossRules (Drools), MySQL, Apache Tomcat, Log4J and Java technologies were essential software for this research project.

Completion of this research project would not have been possible without the relentless support from my family; I thank them for their understanding and support. I also wish to thank my employer for giving adequate time off from office, to do my study work. Finally, my thanks go to my colleagues, friends and all the others who helped.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

DBMS	Database Management System
DSL	Domain Specific Language
HMM	Hidden Markov Model
HTML	Hyper Text Mark-up Language
IPA	International Phonetic Alphabet
IT	Information Technology
JDBC	Java Database Connectivity
JSP	Java Server Pages
JSTL	Java Server Pages Standard Tag Library
LTRL	Language Technology Research Laboratory
OOV	Out of Vocabulary
POC	Proof of Concept
SDK	Standard Development Kit
UCSC	University of Colombo School of Computing
UI	User Interface
UTF	Unicode Transformation Format
XML	eXtensible Mark-up Language
ZWJ	Zero Width Joiner